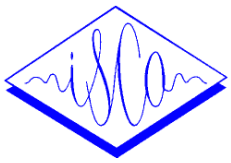

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

STRATEGIES FOR VOICE SEPARATION BASED ON HARMONICITY

Alain de Cheveigné

Laboratoire de Linguistique Formelle, CNRS/Université Paris 7,
case 7003, 2 place Jussieu, 75251, France.

ABSTRACT

A harmonic sound has a spectrum made of discrete components forming a harmonic series, and its wave form is periodic. Voiced speech is approximately harmonic, and this cue appears to be exploited by the auditory system for the perception of speech in a noisy background. This paper addresses the following question: does the auditory system exploit the harmonic structure of a target to segregate it from the background, or else that of a harmonic background to eliminate it.

The issue was considered from two angles. First of all, both strategies were implemented to reduce voiced interference in a speech recognition experiment. The aim was to determine whether one "works better" than the other in this task; the results showed that harmonic cancellation was more effective. In a second experiment human listeners were presented with pairs of vowels, each of which was either harmonic or inharmonic, and requested to identify them both. The responses were scored according to whether the target vowel was harmonic, and whether the other (interfering) vowel was harmonic. It appears that the auditory system exploits harmonicity of the ground (cancellation strategy), a result that is coherent with the outcome of the first experiment.

1. HARMONIC ENHANCEMENT VS CANCELLATION

The regular structure of the spectrum of a harmonic target can be exploited to select its components among those of interfering sounds. We refer to this strategy as *harmonic enhancement*. In a similar fashion the components of a harmonic background may be removed. We refer to this as *harmonic cancellation*. If both target and ground are voiced, both strategies are a priori possible, but their properties and implementation are quite different. It is of theoretical and practical interest to know if one strategy is more effective than the other, and which one (possibly both) is used by the auditory system.

Both strategies have their advantages and drawbacks of principle:

• *Enhancement requires the target to be harmonic but works with any kind of interference.*	• *Cancellation requires the interference to be harmonic but works with any kind of target.*
• *Enhancement requires the fundamental frequency (F0) of the target to be determined. Relatively easy if the Signal-to-Noise Ratio (SNR) is large.*	• *Cancellation requires the F0 of the interference to be estimated. Relatively easy if the SNR is small.*
• *Enhancement causes no spectral distortion of the target, if the target is perfectly harmonic.*	• *Cancellation removes all target components that belong to the harmonic series of the interference.*
• *Effective enhancement requires a filter with a long impulse response.*	• *A filter with a short impulse response can implement cancellation that is perfect if the interference is perfectly harmonic.*

The last point is worth developing. The target enhancement ratio of a harmonic enhancement filter can be defined as:

$$\alpha = f_0 \sum_{k=-\infty}^{+\infty} |H(kf_0)|^2 \Big/ \int_{-\infty}^{\infty} |H(f)|^2 df$$

where H is the transfer function and f_0 is the fundamental frequency of the target. It can be shown that this ratio depends entirely on the values of the autocorrelation function of the filter impulse response sampled at multiples of the target period. In the case of a filter with an impulse response made up of equal pulses spaced at period multiples, the ratio is equal to the number of these pulses [1]. A large enhancement ratio thus requires a long impulse response but the filter will only be effective if the target is stationary over a period of similar length. Voiced speech is stationary over limited periods of time, so enhancement might be less effective than expected. To what point this is true and how the various other trade-offs add up in practice is difficult to predict. For this reason we designed a experiment to compare the effectiveness of the two strategies in a practical task.

2. VOICE SEGREGATION FOR SPEECH RECOGNITION

Both strategies were implemented in an interference-reduction stage for a speech recognition system (Fig. 1).

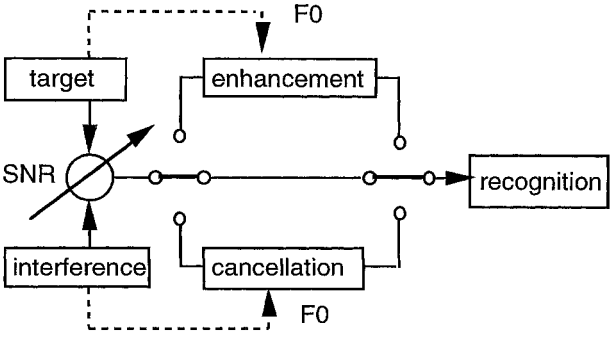

Fig. 1. Schema of experiment 1.

Target speech was mixed with interfering speech at an adjustable signal-to-noise ratio and fed to a speech recognition system. Optionally, the mixture could first be directed through either of two processing stages, each of which implemented one of the strategies of interest. Recognition rate constituted a convenient measure of the effectiveness of processing.

2.1. Experimental details

The task was to recognize short Japanese words among a set of 100 taken from the ATR database. Speech data were sampled at 12 kHz with 12 bit resolution. Recognition was performed by DTW pattern matching using Euclidean distance between strings of 16-coefficient linear spectrum feature vectors calculated from 256-sample Hanning-shaped windows at a 128-sample frame rate.

The F0 estimates necessary for processing were derived from the target and interference speech *before mixing*. The difficult question of how to estimate both F0s from the mixture has been addressed in a previous paper [1]. Cancellation and enhancement were performed in the time domain using filters with impulse responses:

$$h_n = \delta_n - \delta_{n-l} \quad ; \qquad h_n = \frac{1}{K}\sum_{i=0}^{K-1} \delta_{n-il}$$

respectively, where l represents a time lag, δ_n represents the unit impulse, and K is the number of "prongs" in the enhancement filter impulse response.

The same words were used for target tokens and for reference templates; the task was therefore trivial when the target speech was isolated. The same set was also used as interference (to each target word was added some other word). The task was thus rather difficult in the presence of interference. Another consequence, useful for our purpose, was to eliminate any bias in favor of either strategy that might otherwise have occurred due to differences in F0 estimation quality between target and interference. Interference and target were each voiced for 56% of all frames, and simultaneously for 41%. Further details may be found in [2, 3].

2.2. Results using cancellation

The recognition rate of speech mixed with interference is plotted as a function of SNR as the lower curve in Fig. 2. Recognition is perfect in the absence of interference (leftmost point), and impossible in the absence of target (rightmost point). The continuous line in Fig. 2 represents the rates after interference cancellation. Comparison between the two curves shows that cancellation improves recognition at low SNR but not at high SNR. In no case is it sufficient to eliminate the effects of noise and bring the recognition rate near 100% (except for SNR=∞).

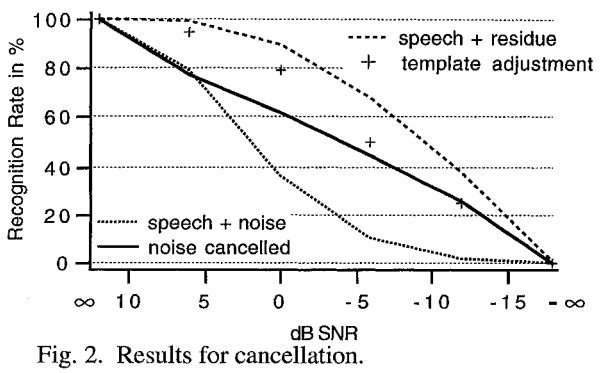

Fig. 2. Results for cancellation.

Two reasons can be advanced to explain this limited effectiveness. The first is that interfering speech is not perfectly harmonic and therefore not perfectly cancelled. The second is that cancellation causes spectral distortion of the target. To separate the contributions of these two factors, the experiment was modified by swapping the linear stages of summation and filtering in Fig. 1. Cancellation was then applied to the interference only instead of both target and interference as before. The rates obtained this way (upper dotted curve of Fig. 2) reflect the contribution of the cancellation residual only. It is great at small SNR, as is to be expected, but at large SNR spectral distortion is the main factor that limits performance (compare the upper dotted line with the continuous line).

2.3. Template adjustment

Spectral distortion causes a mismatch between targets and undistorted templates. To solve this problem, signal restoration techniques have been proposed in which missing information is guessed based on continuity with parts adjacent in time or frequency [4,5,6]. A different approach was followed here: rather than attempt to restore the target we instead applied similar distortion to the reference templates. The distortion caused to the target by the cancellation comb filter can be modelled as a filter with transfer function:

$$|H(f)| = |\sin(2\pi f|T - L|)|$$

where T is the period of the target signal and L is the filter lag parameter [2]. Applying this distortion to the templates results in a clear improvement of recognition at high SNR (crosses in Fig 1). Recognition rate comes closer to 100%, as might be required by a real system. Overall the improvement is roughly equivalent to a 6-9 dB reduction in interference level.

Contrary to simple cancellation which uses only the F0 of the ground, template adjustment requires additional knowledge of the F0 of the target.

2.4. Results using enhancement

Fig. 3 shows results for harmonic enhancement. Recognition rate is plotted as a function of the parameter K (number of "prongs" in the impulse response). For the lowest three SNRs the best value of K seems to be 3; beyond this value recognition rate deteriorates. Apparently the target speech is not sufficiently stationary for longer impulse responses to be effective. The rates obtained with enhancement are in any case less good than those obtained with cancellation (leftmost in the figure).

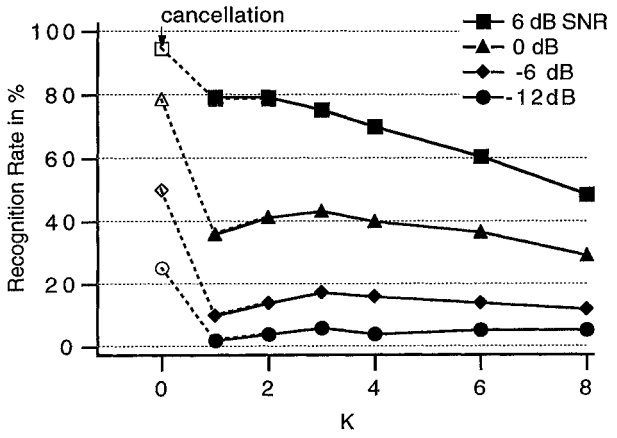

Fig. 3 Results for enhancement. K=1 is equivalent to no processing. Leftmost points are for cancellation.

3. VOICE SEGREGATION IN THE AUDITORY SYSTEM

The second experiment sought to determine which strategy is used by the auditory system. Listeners were presented with pairs of synthetic vowels, each of which was either harmonic or inharmonic (with partials randomly shifted from a harmonic series). The vowels had either the same F0 (125 Hz), or F0s differing by about 3% (the F0 of an inharmonic vowel is defined as the F0 of the harmonic vowel before shifting). Subjects were requested to identify both vowels. Identification of each vowel ("the target") was scored according to its harmonic state and that of the vowel that accompanied it ("the ground").

Vowels were allophones (10 each) of five French vowels: /a/, /e/, /i/, /o/, /u/. The stimuli were presented at a level of 60 dBA via headphones. Stimulus duration was 200 ms including 25 ms onset and offset ramps. Thirty subjects participated in the experiment and responded to 960 stimuli each. Further details may be found in [7,8]

The expected outcome of the experiment is that identification should be better for harmonic than for inharmonic targets if the auditory system uses enhancement. It should be better for harmonic than for inharmonic grounds if it uses cancellation. A priori one expects both strategies to be used whenever possible.

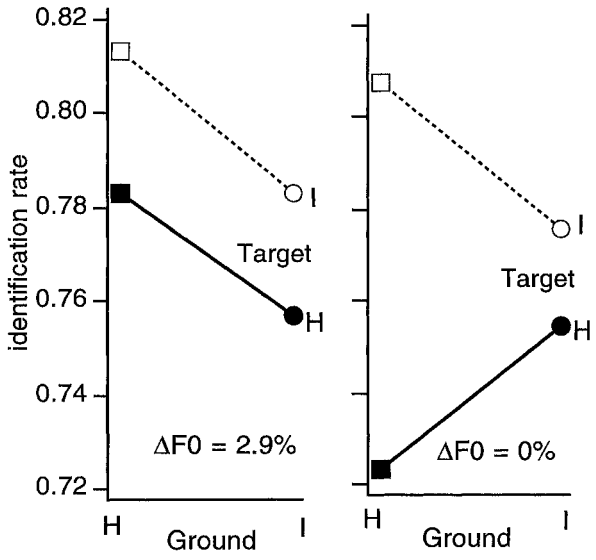

Fig. 4. Identification rate as a function of ground harmonicity for harmonic targets (closed symbols) and inharmonic targets (open symbols) at two values of ΔF0.

Results are plotted in Fig. 4. Considering first the left-hand plot, we see that whatever the nature of the target, identification is better for a harmonic ground. This is consistent with the cancellation hypothesis.

We also see that whatever the nature of the ground, identification is better for an inharmonic rather than harmonic target. This result is unexpected, as it is opposite the trend predicted by the enhancement hypothesis. A possible explanation is that cancellation occurs systematically even when the target is harmonic. Identification of harmonic targets would thus be less good. Similar results are obtained for ΔF0=0% (right-hand plot), except when target and ground are both harmonic in which case identification is poor. This was to be expected, as harmonic segregation is impossible when the harmonic structures of target and ground are the same. The difference between this condition and the corresponding condition at ΔF0=2.9% is coherent with previous results using harmonic synthetic vowels [9,10,11,12,13]

A word of caution: we cannot rule out a different interpretation of our results. All vowels were synthesized in sine phase, but inharmonic vowels may be interpreted as moving into random phase. If the phase of a vowel affects segregation, then the effects we have attributed to harmonicity may be partly the result of phase. This possibility is currently under investigation.

4. DISCUSSION

Both experiments show that harmonic cancellation is an effective strategy. The second suggests in addition that it may be employed by the auditory system to identify vowels within pairs. Recent experiments [14, 15] had reached the same conclusion.

Evidence in favor of enhancement is much weaker. In the first experiment enhancement brought little improvement to recognition rates, even at large values of the enhancement ratio K. The fact that larger values, which entail longer impulse responses, gave smaller recognition rates suggests that the non-stationary nature of speech prevents enhancement from being effective. One must consider however that the database contained F0 values that were narrowly distributed, and therefore most F0 differences were small (less than 15 %). This may have made enhancement particularly difficult in our task, although it did not prevent effective cancellation.

The second experiment found no evidence that the auditory system uses enhancement. This result is surprising as it contradicts the accepted notion that harmonicity labels components of a sound as "belonging together", and thus makes them easier to extract from interference. Many methods and models that have been proposed that are capable of exploiting target harmonicity (see [1] for a review), and it is surprising that the auditory system does not make some use of this cue. It may be that the design of our task somehow prevented evidence of enhancement from showing up in the results. Further experiments are necessary before a definitive conclusion may be drawn.

In the first experiment an estimate of the spectral distortion caused to the target by interference reduction was used to adjust the templates and improve recognition. Distortion of both target and template before matching can be interpreted as a form of non-uniform weighting that puts less emphasis on information that is less reliable (in this case, spectral information that falls on or near the harmonic series of the interference).

Non-uniform weighting constitutes an alternative to spectrum restoration techniques. It might also be of use to interpret "phonemic restoration" effects in speech perception. For example the presence of a burst of noise that masks (or replaces) a speech segment may cause the auditory system to put "zero weight" on that segment in subsequent processing stages, so that the final outcome of perception is governed entirely by the unmasked context as if the segment were replaced by a "wild card". The segment appears "perceptually restored" simply because any other interpretation is unlikely given the lack of *negative* evidence. Bregman [16, p. 356] remarks that restoration occurs if neural activity in all channels is at least as large as might have occurred had the segment been present. If activity is smaller, as if silence is inserted instead of noise, this constitutes strong evidence *against* the presence of the

segment. The difference with a "phoneme restoration mechanism" is of course that no detailed reconstruction of the signal or its correlates need occur.

5. ACKNOWLEDGEMENTS

The first experiment was carried out at ATR Human Information Processing Laboratories in collaboration with Hideki Kawahara, Kiyoaki Aikawa, and Andrew Lea. The second was carried out at IRCAM and LLF in collaboration with Stephen McAdams (LPE and IRCAM) and Jean Laroche and Muriel Rosenberg (ENST). This work has received support from the "Cognitive Sciences" program of the French Ministry of Research and Space.

6. REFERENCES

[1] de Cheveigné, A. (1993), "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing.", JASA 93, 3271-3290.

[2] de Cheveigné, A. (1993), "Time-domain comb filtering for speech separation", ATR Human Information Processing Laboratories technical report TR-H-016.

[3] de Cheveigné, A., H. Kawahara, K. Aikawa and A. Lea (1994), "Speech separation for speech recognition", Journal de Physique IV, C5-545-548.

[4] Parsons, T. W. (1976), "Separation of speech from interfering speech by means of harmonic selection", JASA 60, 911-918.

[5] Weintraub, M. (1985), "A theory and computational model of auditory monaural sound separation", Doctoral dissertation, Stanford University.

[6] Cooke, M. P. and G. J. Brown (1993), "Computational auditory scene analysis: exploiting principles of perceived continuity.", Speech Communication 13, 391-399.

[7] de Cheveigné, A., S. McAdams, J. Laroche and M. Rosenberg (1994), "Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement", in preparation.

[8] de Cheveigné, A., S. McAdams, J. Laroche and M. Rosenberg (1994), "Identification de voyelles simultanées harmoniques et inharmoniques", Journal de Physique IV, C5-553-556.

[9] Assmann, P. F. and Q. Summerfield (1990), "Modeling the perception of concurrent vowels: vowels with different fundamental frequencies", JASA 88, 680-697.

[10] Culling, J. F. and C. J. Darwin (1993), "Perceptual separation of simultaneous vowels: within and across-formant grouping by F0", JASA 93, 3454-3467.

[11] Scheffers, M. T. M. (1983), "Sifting vowels", Doctoral dissertation, University of Gröningen.

[12] Summerfield, Q. and P. F. Assmann (1991), "Perception of concurrent vowels: effects of harmonic misalignment and pitch-period asynchrony", JASA 89, 1364-1377.

[13] Zwicker, U. T. (1984), "Auditory recognition of diotic and dichotic vowel pairs", Speech Communication 3, 256-277.

[14] Summerfield, Q. and J. F. Culling (1992), "Periodicity of maskers not targets determines ease of perceptual segregation using differences in fundamental frequency", 124th meeting of the ASA

[15] Lea, A. (1992), "Auditory models of vowel perception", Doctoral dissertation, University of Nottingham.

[16] Bregman, A. S. (1990), Auditory scene analysis, MIT Press: Cambridge, Mass.